Statistica Sinica **5**(1995), 333-349

ROBUST ESTIMATION FOR SEMIPARAMETRIC EXPONENTIAL MIXTURE MODELS

Larry Zaiqian Shen

The Procter and Gamble Company

Abstract. B-optimal robust estimates are considered for semiparametric exponential mixture models, under the perception that the data may have been contaminated. The B-optimal robust influence functions are defined by Hampel's variational problem: minimizing the asymptotic variances over the class of influence functions bounded by a constant. Explicit B-optimal influence functions are calculated for the semiparametric exponential mixture models. The one-step procedure is used to construct the B-optimal robust estimates from the B-optimal influence functions. A small Monte-Carlo study is conducted for the semiparametric two-sample exponential mixture model to confirm the theory.

Key words and phrases: B-optimal, bounded influence function, exponential mixture model, Hampel's problem, most robust estimate.

1. Introduction

Let $(\mathbf{X}, \mathcal{B}, \mu)$ be a sample space, where $\mathcal{B}$ is a Borel field and μ is the Lebesgue measure. Consider a semiparametric mixture model on $(\mathbf{X}, \mathcal{B}, \mu)$ of the form

$$P = \{P_{\theta,G} : \theta \in \Theta \subset R^d, G \in \mathbf{G}\} \tag{1}$$

with $P_{\theta,G} = \int Q_{\theta,\eta} dG(\eta)$, where $Q = \{Q_{\theta,\eta} : \theta \in \Theta \subset R^d, \eta \in H \subset R^q\}$ is a regular parametric model and G is the distribution of η. Each member $Q_{\theta,\eta}$ of Q has density $f(\cdot, \theta, \eta)$. The set $\mathbf{G}$ contains distribution functions on H. The unconditional density corresponding to $P_{\theta,G}$ is denoted by $f(\cdot, \theta, G)$.

Motivating this model are situations when estimating the parameter θ in the parametric model Q, we bring in another *incidental* parameter η_j with each sampling X_j. The number of parameters becomes large as the sample size increases. This poses difficulty for consistent estimation of θ. Neyman and Scott (1948) were the first to notice such a phenomenon. Instead of trying to estimate θ and the parameters η_j simultaneously, we treat the η_js as nuisance parameters which come from an unknown distribution G. Thus it is possible to reduce the number of parameters to be estimated. The following examples are exponential mixture models, which will be studied in this paper.

Example 1. (The Neyman-Scott Models) Assume that the components of a random vector $X = (X^{(1)}, \ldots, X^{(k)})^T$ are i.i.d. samples from $N(\nu, \sigma^2)$.

Model L. (Location) In Example 1, suppose the main interest is to estimate ν, the location of X, then the main parameter is $\theta = \nu$. The incidental parameter is $\eta = -(2\sigma^2)^{-1}$ which is assumed to follow an unknown distribution G. Thus, for each given η, the random vector X has density

$$f(x, \theta, \eta) = \exp\left\{\eta \sum_{j=1}^{k} (x^{(j)} - \theta)^2 - \frac{k}{2} \ln(-\frac{\pi}{\eta})\right\}. \tag{2}$$

The unconditional density of X is $\int f(x, \theta, \eta) G(d\eta)$.

Model S. (Scale) In Example 1, one may choose to estimate the variance σ^2 instead, with the location ν as the incidental parameter. Then the main parameter is $\theta = -(2\sigma^2)^{-1}$; and the nuisance parameter is $\eta = \nu/\sigma^2$. For model S, the conditional density of X given η is

$$f(x, \theta, \eta) = \exp\left(\eta \sum_{j=1}^{k} x^{(j)} + \theta \sum_{j=1}^{k} (x^{(j)})^2 - \frac{k}{2}\left\{-\frac{\eta^2}{2\theta} + \ln(-\frac{\pi}{\theta})\right\}\right).$$

Again, assume that η has a distribution $G(\cdot)$. The class of (unconditional) distributions of X forms a mixture model. Neyman and Scott (1948) introduced this model and pointed out that as the number of parameters becomes large, the maximum likelihood estimate of σ^2 is no longer consistent.

Example 2. (Two Sample Exponential Mixture Model) Consider a bivariate random vector $X = (x^{(1)}, x^{(2)})^T$ which, for given $\eta > 0$, has a density

$$f(x^{(1)}, x^{(2)}, \theta, \eta) = \exp\{-\eta(x^{(1)} + \theta x^{(2)}) + \ln(\eta^2\theta)\}, \ \ x_1, x_2 > 0.$$

The nuisance parameter η is distributed according to $G(\cdot)$.

Here $x^{(1)}$ and $x^{(2)}$ are independent and have exponential distributions with rates η and $\theta\eta$, respectively. The ratio θ between the two rates is the main parameter; and η, the baseline rate of $x^{(1)}$, is the nuisance parameter which has unknown distribution G.

All the above examples share a common structure: the density $f(x, \theta, \eta)$ is exponential in η. These mixture models are called *semiparametric exponential mixture models* and have been considered by many authors. Lindsay (1983) studied efficient score functions for exponential mixture models. Van der Vaart (1988) constructed efficient estimates for more general mixture models. A more complete theory regarding efficient estimation in mixture models may be found in

Bickel, Klaassen, Ritov, and Wellner (1993) ([BKRW] hereafter). We are going to follow their approach in this paper. According to [BKRW], a general exponential mixture model has the form: for given η,

$$f(x,\theta,\eta) = \exp\{\eta^T T(x,\theta) + S(x,\theta) - b(\theta,\eta)\}. \tag{3}$$

This model has two special cases:

$$f(x,\theta,\eta) = \exp\{\eta^T T(x,\theta) - b(\theta,\eta)\} \tag{4}$$

and

$$f(x,\theta,\eta) = \exp\{\eta^T T(x) + \theta^T S(X) - b(\theta,\eta)\}. \tag{5}$$

It can be seen that Model L in Example 1 and Example 2 belong to type (4) and Model S belongs to type (5). Only exponential mixture models will be considered in this paper.

Let $X_1,\ldots,X_n$ be i.i.d. samples from $P_{\theta,G}$. For simplicity, assume that θ is of one-dimension. We shall focus on asymptotically linear estimates of θ, which can be written as

$$\hat{\theta}_n = \theta + n^{-1}\sum_{i=1}^{n}\psi(X_i;\theta,G) + o_{\theta,G}(n^{-\frac{1}{2}}),$$

where ψ is called the *influence function* of $\hat{\theta}_n$. The asymptotic variance of $\hat{\theta}_n$ is given by $V(\psi) = \int \psi^2 dP_{\theta,G}$. An influence function is called the *efficient influence function* if it minimizes $V(\psi)$ among all influence functions. The corresponding estimate will be called *efficient estimate.*

Efficient influence functions are unbounded for many models, which means that a single outlier in the data may have large influence. When the presence of outliers is suspected it is preferable to use (*B-robust*) estimates with bounded influence functions. However, B-robust estimates may be less efficient when the data actually contain no contamination. To reduce the cost of using robust estimates, we consider an influence function that solves Hampel's variational problem: minimizing the asymptotic variance $V(\psi)$ among all influence functions bounded by a constant C (e.g., Hampel et al. (1986)). The influence function solving Hampel's problem is called *B-optimal.* In a different context, Wu (1990) also studies robust estimation problem in semiparametric mixture models and makes minor compromise between efficiency and robustness.

This paper is arranged in the following way: Section 2 gives explicit form of the B-optimal influence functions in exponential mixture models; Section 3 constructs the optimal estimates from the optimal influence functions; Section

4 applies the general theory to the above examples, and Section 5 presents the result from a small simulation study.

2. B-Optimal Influence Functions

We calculate the optimal influence functions in this section. The same notation as that in Section 4.5 of [BKRW] will be adopted.

For a semiparametric density $f(x,\theta,G)$, define the *score function* $\dot{\mathbf{l}}(x,\theta,G)$ as the partial derivative of $\log\{f(x,\theta,G)\}$ with respect to θ. The tangent space is defined as the L_2-closure of the linear span of the set:

$$\{\frac{\partial}{\partial\beta}\log f(x,\theta,G_\beta) : \{G_\beta : 0 \le \beta \le 1\} \text{ is a parametric subset of } \mathbf{G}\}.$$

Conventions

1. The density g and the distribution G of η are used interchangeably.
2. Write $\dot{S}(x,\theta)$, $\dot{T}(x,\theta)$, and $\dot{b}(\theta,\eta)$ for the partial derivatives of $S(x,\theta)$, $T(x,\theta)$, and $b(\theta,\eta)$ with respect to θ.
3. Let S_θ and T_θ stand for $S(\theta,x)$ and $T(\theta,x)$ respectively.

According to Theorem 4.5.1 and Corollary 4.5.1 of [BKRW], the score function for θ is

$$\dot{\mathbf{l}}_1(X,\theta,g) = \dot{T}(X,\theta)E(\eta|T) + \dot{S}(X,\theta) - E\{\dot{b}(\theta,\eta)|T\}, \tag{6}$$

and the tangent space for G is $\dot{\mathbf{P}}_2 = \{w(T_\theta) : w(T_\theta) \in L_2(P),\ \text{and } Ew(T_\theta) = 0\}$. Thus the *efficient score function* for estimating θ is given by

$$\dot{\mathbf{l}}_1^* = E(\eta|T)\{\dot{T}(X,\theta) - E(\dot{T}(X,\theta)|T)\} + \dot{S}(X,\theta) - E\{\dot{S}(X,\theta)|T\}.$$

The efficient influence function is $\tilde{\mathbf{l}} = \dot{\mathbf{l}}_1^*/\|\dot{\mathbf{l}}_1^*\|^2$, which is unbounded if either $T(x,\theta)$ or $S(x,\theta)$ is unbounded. An intuitive way to achieve boundedness is to truncate the efficient score function wherever it is too large. However, the resulting function may fail to be an influence function. Some correction has to be made to overcome this.

To explain the idea formally, let $h_c(x) = \min(1, c/|x|)x$ be the Huber truncating function. The optimal influence function ψ is then expected to be of the form

$$\psi = h_c(\lambda_0\dot{\mathbf{l}}_1 + w_0(T_\theta) + a_0), \tag{7}$$

where $w_0(T_\theta) \in \dot{P}_2$; λ_0 and a_0 are constants. According to [BKRW], in order for ψ to be an influence function it has to satisfy the following conditions.

Consistency:

$$\int \psi \; dP_{\theta,G} = 0, \tag{i}$$

$$\int \psi \; \dot{l}_1 \; dP_{\theta,G} = 1, \tag{ii}$$

$$\int \psi \; w(T_\theta) \; dP_{\theta,G} = 0 \;, \qquad \text{for any} \;\; w(T_\theta) \in \dot{P}_2. \tag{iii}$$

Lemma 1. *If an influence function ψ of the form* (7) *satisfies conditions* (i)-(iii) *then ψ is a B-optimal influence function.*

Proof. The proof is similar to that of Theorem 4.1 of Hampel et al. (1986).

The conditional expectation of X given T plays an important role in the calculation of the efficient influence function. Write $d\lambda(x,\theta,t) = \mu(x|T_\theta = t)$ for the conditional measure of μ given $T_\theta = t$. Define

$$f(x,\theta|t) = e^{S(x,\theta)}/E_\mu(e^{S(x,\theta)}|T_\theta = t), \qquad \text{a.e. } x, \; d\lambda(x,\theta,t).$$

Then $f(x,\theta|t)$ is the conditional density of X given $T_\theta = t$ with respect to $d\lambda(x,\theta,t)$. Note that $f(x,\theta|t)$ does not depend on G. This suggests that we should focus the attention on $f(x,\theta|t)$. Define for $c > 0$ a function $\rho_c(\cdot) : R \to R^+$ by

$$\rho_c(x) = \begin{cases} x^2/2, & \text{if } |x| \le c, \\ c^2/2 + c(|x| - c), & \text{if } |x| > c. \end{cases}$$

It is easy to see that $\rho_c'(x) = h_c(x)$.

Theorem 1. *For any $c_0 > 0$, let $w(t,\theta,G)$ minimize the following expression:*

$$E\left(\rho_{c_0}\{\dot{l}_1(x,\theta,G) + w\}\Big| \, T_\theta = t\right). \tag{8}$$

Introduce $\psi_1 = h_{c_0}\{\dot{l}_1(x,\theta,G) + w(T(x,\theta),\theta,G)\}$ and $\lambda = \int \psi_1 \dot{l}_1(x,\theta,G) dP_{\theta,G}$. Then $\psi_0 = \lambda^{-1}\psi_1$ is the B-optimal influence function with bound $c = \lambda^{-1}c_0$.

Proof. The existence of $w(\cdot)$ follows from convexity of the function $\rho_c(\cdot)$. It is evident that ψ_0 can be written in the form (7) with $\lambda_0 = \lambda^{-1}$, $a_0 = \lambda^{-1}Ew(T_\theta(x),\theta,G)$, and $w_0(T_\theta) = \lambda^{-1}\{w(T_\theta) - Ew(T_\theta)\}$. Since $w(t,\theta,G)$ satisfies $E(h_{c_0}\{\dot{l}_1(x,\theta,G) + w\}|T_\theta = t) = 0$, it follows that $E(\psi_0|T_\theta = t) = 0$, for any t. Therefore consistency conditions (i) and (iii) are satisfied. Finally condition (ii) follows from the definition of λ.

One point worth mentioning is that the term $E\{\dot{b}(\theta,\eta)|T_\theta\}$ which appears in the expression of the score function may be ignored because it belongs to the space $\{1\} \oplus \dot{P}_2$ and can be assimilated into $w(T)$.

Denote by $p_T(t,\theta,G)$ the marginal density of T_θ with respect to Lebesgue measure; then, for some function $\pi(t,\theta)$,

$$p_T(t,\theta,G) = \int \exp\{\eta^T t - b(\theta,\eta)\}\ dG(\eta)\pi(t,\theta).$$

Since $\dot{\mathbf{l}}_1$ depends on G only through $E(\eta|T_\theta)$, one can show that $\dot{\mathbf{l}}_1$ depends on G only through $p_T(\cdot)$ according to the following relation:

$$E(\eta|T=t,\theta) = \frac{\nabla_t p_T}{p_T}(t,\theta,G) - \frac{\nabla_t \pi}{\pi}(t,\theta). \tag{9}$$

where $\nabla_t p_T$ is the gradient of $p_T(\cdot)$ with respect to t.

In particular when the model is given by (5), we have $T(x,\theta) = T(x)$. Then the score function equals $\dot{\mathbf{l}}_1(x,\theta) = S(x) - E\{\dot{b}(\theta,\eta)|T\}$; and the optimal influence function becomes $\psi(x) = \lambda^{-1}h_c\{S(x)+w(T(x),\theta)\}$, where $w(t,\theta)$ solves equation $E(h_c\{S(x)+w\}|T=t)=0$. In this situation, the problem is reduced to robust estimation for a parametric model derived by conditioning X on T. Since G is no longer involved, the calculation of the optimal influence function becomes much easier.

3. Construction of the Optimal Robust Estimates

In this section we shall apply Klaassen's method (Klaassen (1987)) to construct an asymptotically linear estimate of θ corresponding to the optimal influence function. Let $X_1,\ldots,X_n$ be i.i.d. random samples from $P_{\theta,G}$. Klaassen's procedure will consist of the following steps.

First, we need to find a preliminary estimate $\tilde{\theta}_n$ that is $\sqrt{n}$-consistent. We assume, temporarily, the existence of $\tilde{\theta}_n$. In Section 4 we provide initial estimates for some examples. Using either the method of discretization introduced by LeCam (1956) or the sample splitting method introduced by Bickel (1982), we may treat $\tilde{\theta}_n$ as nonrandom and hence write it as θ_n.

Secondly, we would like to estimate the score function. Assume that θ_n is a discretized preliminary estimate. Define $T_i = T(X_i,\theta_n)$, $i=1,\ldots,n$. Write $\phi(x)$ for the density function of the standard normal distribution. Let $\sigma_n > 0$ be a bandwidth tending to zero at a certain rate. For given t and θ, the kernel estimate of $p_T(t,\theta,G)$ is

$$\hat{p}_n(t,\theta) = \frac{1}{n\sigma_n}\sum_{i=1}^{n}\phi\left(\frac{t-T(X_i,\theta)}{\sigma_n}\right).$$

Let the partial derivative of $\hat{p}_n(t,\theta)$ with respect to t be denoted by $\hat{p}'_n(t,\theta)$. Following Bickel (1982), for given σ_n, c_n, d_n, $e_n > 0$ define

$$\hat{q}_n(t,\theta) = \begin{cases} (\hat{p}'_n/\hat{p}_n)(t,\theta), & \text{if } |\hat{p}_n(t,\theta)| \geq d_n,\ |t| \leq e_n,\ \ |(\hat{p}'_n/\hat{p}_n)(t,\theta)| \leq c_n, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Since θ_n is close to θ, $\hat{p}_n(t,\theta_n)$ and $\hat{q}_n(t,\theta_n)$ may serve as estimates of $p_T(t,\theta,G)$ and $q_T(t,\theta,G)=p'_T(t,\theta,G)/p_T(t,\theta,G)$, respectively. The score function $\dot{\mathbf{l}}_1(x,\theta,G)$ may be estimated by

$$\hat{\mathbf{l}}_n(x,\theta_n) = \dot{T}(x,\theta_n)\hat{E}\{\eta|T(x,\theta_n),\theta_n\} + \dot{S}(x,\theta_n) - E\{\dot{b}(\theta_n,\eta)|T(x,\theta_n)\}, \quad (11)$$

where by (9), $\hat{E}_n(\eta|T,\theta_n) = \hat{q}_n(T,\theta_n) - (\pi'/\pi)(T,\theta_n)$. Let $\hat{w}_n(t,\theta_n)$ minimize

$$E\left(\rho_c\left\{\hat{\mathbf{l}}_n(x,\theta_n)+w\right\}\Big| T_{\theta_n}=t\right).$$

Then $\hat{w}(t,\theta_n)$ solves the equation $E(h_c\{\hat{\mathbf{l}}_n(x,\theta_n)+w\}|T_{\theta_n}=t)=0$. The estimate of ψ_1 defined in Theorem 1 is given by

$$\hat{\psi}_n(x,\theta_n;X_1,\ldots,X_n) = h_c\left\{\hat{\mathbf{l}}_n(x,\theta_n)+\hat{w}_n(T(x,\theta_n),\theta_n)\right\}. \quad (12)$$

It remains to estimate λ, the inner product between ψ_1 and $\dot{\mathbf{l}}_1$. For the sake of later proofs, we usually split the samples, though we may not have to do so in practical situations. Following Klaassen's approach, let $m_n \approx a_0 n$ with a_0 a constant. Split the samples $\{X_1,\ldots,X_n\}$ into $\{X_1,\ldots,X_{m_n}\}$ and $\{X_{m_n+1},\ldots,X_n\}$. The first part $\{X_1,\ldots,X_{m_n}\}$ is used to obtain the estimates $\hat{\psi}_n$ and $\hat{\mathbf{l}}_n$ as before. We estimate λ from the second part by

$$\hat{\lambda}_n = \frac{1}{n-m_n}\sum_{i=m_n+1}^{n}\hat{\psi}_n(X_i,\theta_n)\hat{\mathbf{l}}_n(X_i,\theta_n). \quad (13)$$

The optimal influence function can be estimated by

$$\hat{\psi}_0(x,\theta_n;X_1,\ldots,X_n) = \hat{\lambda}_n^{-1}\hat{\psi}_n(x,\theta_n;X_1,\ldots,X_n). \quad (14)$$

Having estimated the optimal influence function, we calculate the one-step estimate:

$$\hat{\theta}_n = \theta_n + \frac{1}{n-m_n}\sum_{i=m_n+1}^{n}\hat{\psi}_0(X_i,\theta_n;X_1,\ldots,X_n). \quad (15)$$

Theorem 2. *Let θ_n be a discretized preliminary $\sqrt{n}$-consistent estimate of θ. Assume $\hat{\lambda}_n$ and $\hat{\psi}_0$ are estimates of λ and ψ, defined by* (13) *and* (14), *respectively. Then the one-step estimate $\hat{\theta}_n$ in* (15) *is an asymptotically linear estimate of θ with the optimal influence function ψ_0 as defined by Theorem* 1.

The proof will use Theorem 2.1 of Klaassen (1987). The key steps are to show that the conditions for the theorem are valid. The following regularity conditions are necessary.

(R.1) Functions $p_T(t,\theta,G)$, $p'_T(t,\theta,G)$, and $\pi(t,\theta)$ are all continuous in (t,θ).

(R.2) For fixed G, $\{p_T(\cdot,\theta,G):\ \theta\in\Theta\}$ is a regular parametric model.

Klaassen's Theorem 2.1 is based on conditions (1.1)-(1.2), (2.1)-(2.2), and (1.4)-(1.5) of Klaassen (1987). To avoid confusion, we mark 'K' in front of the above numbers for Klaassen's conditions. For instance, Klaassen's condition (1.1) will be denoted by (K1.1). Note that the definition of the optimal influence function ψ_0 ensures conditions (K1.1) and (K1.2) and that condition (R.2) implies condition (K2.1). Then it remains to establish (K2.2) for ψ_0 and (K1.4)-(K1.5) for $\hat{\psi}_0$.

To show that $\psi_0(x,\theta,G)$ satisfies (K2.2), we need continuity of $f(x,\theta|t)$ and $\psi_0(x,\theta,G)$ in θ. Difficulties arise since $d\lambda(x,\theta,t)$ is no longer absolutely continuous with respect to the Lebesgue measure on the space X. In order to define the density of $d\lambda(x,\theta,t)$, we have to focus on its support, which resides in a lower dimensional space. In the following we shall expand $f(x,\theta|t)$ and study its smoothness.

Let $R_\theta\subset R$ stand for the range of mapping $T(x,\theta):X\times\Theta\to R$. Write $\underline{x}=(x_1,\underline{x}_2^T)^T$ with $\underline{x}_2\in R^{k-1}$. For each $t\in R_\theta$, define

$$A_\theta(t)=\{\underline{x}_2:\ \text{There exists an } x_1 \text{ such that } T(x_1,\underline{x}_2,\theta)=t\}.$$

Assume that there exist regions $A_1,\ldots,A_m\subset R$, with $A_i\cap A_j=\emptyset$ $(i\neq j)$ and $\bigcup_{i=1}^m A_i=R$, such that $T(x,\theta)=T(x_1,\underline{x}_2,\theta)$ is strictly monotonic in x_1 in the regions $A_1,\ldots,A_m$. Define inverse functions $x_1^i(t,\underline{x}_2,\theta):R_\theta\times R^{d-1}\times\Theta\to A_i$ as satisfying $T(x_1^i(t,\underline{x}_2,\theta),\underline{x}_2,\theta)=t$, for all $i,\underline{x}_2$ and θ. Furthermore, assume that each $x_1^i(t,\underline{x}_2,\theta)$ is jointly continuous in $(t,\underline{x}_2,\theta)$ and has continuous partial derivative with respect to t.

Simple calculation shows that $f(x,\theta|t)$ has support on $A_\theta(t)$. Define

$$p_i(\underline{x}_2,\theta|t)=\frac{1}{p_T(t,\theta,G)}p(x_1^i(t,\underline{x}_2,\theta),\underline{x}_2,\theta)|\frac{\partial}{\partial t}x_1^i(t,\underline{x}_2,\theta)|\ 1_{A_\theta(t)}.$$

Then for any function $g(x)$ we have

$$\begin{aligned}E(g(X)|T=t)&=\int g(x)f(x,\theta|t)d\lambda(x,\theta,t)\\&=\int\cdots\int\sum_{i=1}^m g(x_1^i(t,\underline{x}_2,\theta),\underline{x}_2)p_i(\underline{x}_2,\theta|t)d\underline{x}_2.\end{aligned}\tag{16}$$

We introduce more regularity conditions.

(R.3) The conditional expectations $E(|\dot{T}(X,\theta)|\,|\,T_\theta=t)$ and $E(|\dot{S}(X,\theta)|\,|\,T_\theta=t)$ are continuous in (t,θ) for almost all t.

(R.4) For each i and each $\underline{x}_2$, $p_i(\underline{x}_2, \theta|t)$ is continuous in (t, θ) for almost all t.

Lemma 2. *Assume that conditions* (R.3) *and* (R.4) *are valid. Let* $\psi_0(x, \theta, G)$ *be defined by Theorem* 1. *Then* $\psi_0(x, \theta, G)$ *is continuous in* θ.

Proof. For $(\theta_n, t_n) \to (\theta, t)$, define

$$g_n(x) = \dot{T}(x, \theta_n) \left\{ q(t_n, \theta_n, G) - \frac{\pi'}{\pi}(t_n, \theta_n) \right\} + \dot{S}(x, \theta_n)$$

and

$$g(x, t, \theta, G) = \dot{T}(x, \theta) \left\{ q(t, \theta, G) - \frac{\pi'}{\pi}(t, \theta) \right\} + \dot{S}(x, \theta).$$

It is evident that $g_n(x)$ converges to $g(x, t, \theta, G)$ as n tends to infinity. By (16), conditions (R.3) and (R.4) imply conditions (C.3)-(C.5) in Theorem 3 of Shen (1994). It follows that $h_c\{g_n + w(t_n, \theta_n)\}(x)$ converges to $h_c\{g + w(t, \theta)\}(x)$ for almost all x. Hence the function

$$h_c \left(\dot{T}(x, \theta) \left\{ q(t, \theta, G) - \frac{\pi'}{\pi}(t, \theta) \right\} + \dot{S}(x, \theta) + w(t, \theta, G) \right)$$

is continuous in (t, θ). Since $T(x, \theta)$ is continuous in θ, so is $\psi_0(x, \theta, G)$.

Proposition 1. *Let* $\{p(x, \theta) : \theta \in \Theta \subset R\}$ *be a regular parametric model. Assume that the density function* $p(x, \theta)$ *and the score function* $\dot{\mathbf{l}} = \frac{\partial}{\partial \theta} \log p(x, \theta)$ *are continuous in* θ *for almost every* x. *Suppose a function* $\psi(x, \theta)$ *is bounded, continuous in* θ, *and satisfies* $\int \psi(x, \theta) p(x, \theta) dx = 0$ *and* $\int \psi(x, \theta) \dot{\mathbf{l}} p(x, \theta) dx = 1$ *for every* θ. *Let* $X_1, \ldots, X_n$ *be i.i.d. samples from* $p(x, \theta)$. *Then for any sequence* $\{\theta_n\}$ *with* $\sqrt{n}|\theta_n - \theta| = O(1)$,

$$\sqrt{n} \left(\theta_n - \theta + \frac{1}{n} \sum_{i=1}^{n} \{\psi(X_i, \theta_n) - \psi(X_i, \theta)\} \right) = o_\theta(1). \tag{17}$$

Lemma 2 and Proposition 1 together indicate that ψ_0 satisfies condition (K2.2). The proof of Proposition 1 can be found in Shen (1992). Next, we show that the estimate $\hat{\psi}_0(\cdot)$ defined in (14) satisfies conditions (K1.4) and (K1.5). First it needs to be shown that the $\hat{q}_n(\cdot)$ defined in (10) tends to $q(t, \theta, g)$. Following Bickel (1982), we assume $c_n \to \infty$, $e_n \to 0$ and $d_n \to 0$ such that $\sigma_n c_n \to 0$, and $e_n \sigma_n^{-3} = o(n)$.

Lemma 3. *Assume the above conditions hold and* $\sqrt{n}(\theta_n - \theta) = O(1)$. *Suppose the regularity conditions* (R.1)-(R.4) *hold. Let* $\hat{q}_n(t, \theta)$ *be defined by* (10). *Then for almost every* t, $\hat{q}_n(t, \theta_n)$ *tends to* $q(t, \theta, G)$ *in* $P_{\theta_n, G}$.

Proof. When $\sqrt{n}(\theta_n - \theta) = O(1)$, condition (R.2) implies that

$$\mathcal{L}_{P_{\theta_n, g}}(T(X_1, \theta_n), \ldots, T(X_n, \theta_n)) <> \mathcal{L}_{P_{\theta, g}}(T(X_1, \theta), \ldots, T(X_n, \theta)),$$

where "<>" stands for mutual contiguity between two probability distributions. Since $\hat{q}(t,\theta)$ is determined only by $(T_\theta(X_1),\dots,T_\theta(X_n))$, the lemma holds if and only if $\hat{q}(t,\theta)\to q(t,\theta,G)$ in $P_{\theta,G}$, which is a direct result from (6.12) and (6.13) of Bickel (1982).

This lemma also implies that $\hat{E}(\eta|t,\theta_n)$ tends to $E(\eta|t,\theta)$ in $P_{\theta_n,G}$. Thus, $\hat{\mathbf{l}}_n(x,\theta_n)$ defined by (11) is a consistent estimate of the score function $\dot{\mathbf{l}}_1(x,\theta,G)$.

Lemma 4. *Let $\hat{\psi}_n(x,\theta_n,X_1,\dots,X_n)$ be defined by* (12) *and $\psi_1(x,\theta,G)$ by Theorem* 1. *Assume that conditions* (R.1)-(R.4) *hold. Then*

$$\int|\hat{\psi}_n(x,\theta_n;X_1,\dots,X_n)-\psi_1(x,\theta,G)|^2dP_{\theta_n,G}(x)\to 0\quad in\ P_{\theta_n,G}. \tag{18}$$

Proof. Note that

$$\begin{aligned}&\int|\hat{\psi}_n(x,\theta_n;X_1,\dots,X_n)-\psi_1(x,\theta,G)|^2dP_{\theta_n,G}(x)\\ =&\int\Big(\int|\hat{\psi}_n(x,\theta_n;X_1,\dots,X_n)-\psi_1(x,\theta,G)|^2f(x,\theta_n|t)d\lambda(x,\theta_n,t)\Big)\,p_T(t,\theta_n,G)dt\\ =&\int\!\!\int\Big(h_c\Big\{\dot{T}(x,\theta_n)\hat{E}(\eta|t,\theta_n)+\dot{S}(x,\theta_n)+\hat{w}_n(t,\theta_n)\Big\}\\ &\qquad-h_c\Big\{\dot{T}(x,\theta)E(\eta|t,\theta)+\dot{S}(x,\theta)+w(t,\theta,G)\Big\}\Big)^2\\ &\qquad\times f(x,\theta_n|t)d\lambda(x,\theta_n,t)p_T(t,\theta_n,G)dt.\end{aligned}\tag{19}$$

It follows from Lemma 3 that for fixed t

$$\dot{T}(x,\theta_n)\hat{E}(\eta|t,\theta_n)+\dot{S}(x,\theta_n)\longrightarrow\dot{T}(x,\theta)E(\eta|t,\theta)+\dot{S}(x,\theta)\qquad\text{in } P_{\theta_n,G}.$$

By (R.3), (R.4) and Theorem 3 of Shen (1994) again, we conclude that

$$\begin{aligned}&h_c\Big\{\dot{T}(x,\theta_n)\hat{E}(\eta|t,\theta_n)+\dot{S}(x,\theta_n)+\hat{w}_n(t,\theta_n)\Big\}\\ \longrightarrow\quad&h_c\Big\{\dot{T}(x,\theta)E(\eta|t,\theta)+\dot{S}(x,\theta)+w(t,\theta,G)\Big\}\qquad\text{in } P_{\theta_n,G}.\end{aligned}$$

By the Dominated Convergence Theorem, the inside integral in (19) converges to zero in $P_{\theta_n,G}$. Since the inside integral itself is also bounded by c, the entire integral tends to zero in $P_{\theta_n,G}$ by the Dominated Convergence Theorem and Scheffe's theorem applied to $P_T(\cdot,\theta_n,G)$.

Corollary 1. *Let $\hat{\lambda}_n$ and λ be defined by* (13) *and Theorem* 1, *respectively. Then $\hat{\lambda}_n\longrightarrow\lambda$ in $P_{\theta_n,G}$.*

The proof for the corollary follows along the same lines as in Bickel (1982).

Proof of Theorem 2. It only remains to establish condition (K1.5). However, by definition of $\hat{\psi}_n$,

$$\int \hat{\psi}_n(x,\theta_n;X_1,\ldots,X_n)dP_{\theta_n,G} = \int E\left\{\hat{\psi}_n(x,\theta_n;X_1,\ldots,X_n)\middle| T_{\theta_n}\right\}dP_{\theta_n,G} = 0.$$

4. Examples

In this section we calculate B-optimal influence functions for the examples introduced in Section 1.

The first one we consider is the Neyman-Scott model (S). Recall that for given η, the random vector $X = (X^{(1)},\ldots,X^{(k)})$ has density $f(x,\theta,\eta) = \exp\{\eta T(x) + \theta S(x) - b(\theta,\eta)\}$, where $T(X) = \sum_{j=1}^k X^{(j)}$, $S(X) = \sum_{j=1}^k (X^{(j)})^2$ and $b(\theta,\eta) = (k/2)\{-\eta^2/(2\theta) + \ln(-\pi/\theta)\}$.

Assume that the marginal distribution of η is $G(\cdot)$. According to [BKRW], the score function for θ is $\dot{\mathbf{l}}_1 = S(x) - E\{\dot{b}(\theta,\eta)|T\}$. For $c > 0$, choose $w(T)$ such that

$$E\{h_c(\dot{\mathbf{l}}_1 + w)|T\} = 0. \tag{20}$$

Let $U = -2\theta\sum_{j=1}^k (X^{(j)} - T(X)/k)^2$; then $S(X) = -(2\theta)^{-1}U + k^{-1}T^2(X)$. It is well known that U is independent of $T(X)$ and has a χ^2_{k-1} distribution. Write $g_{k-1}(u)$ for the χ^2_{k-1} distribution density. Then (20) is equivalent to

$$\int h_c(-\frac{1}{2\theta}u + w_1)g_{k-1}(u)du = 0.$$

The function $w_1(\cdot)$ depends only on θ. Since $g_{k-1}(\cdot)$ is a non-atomic density, it is easy to show that $w_1(\theta)$ is a continuous and strictly monotonic function of θ. Define $\lambda(\theta) = \int h_c(-U/2\theta + w_1)S(X)dP_{\theta,G}$. By Theorem 1, $\psi_0 = \lambda^{-1}h_c(-(2\theta)^{-1}U + w_1)$ is an optimal influence function corresponding to bound $\lambda^{-1}c$.

Next we construct the optimal estimate corresponding to the optimal influence function just derived. Suppose we have i.i.d. samples $X_i = (X_i^{(1)},\ldots,X_i^{(k)})$, $i = 1,\ldots,n$, from the Neyman-Scott model. Let $\hat{\theta}_n$ be the M-estimate solving the equation

$$\sum_{i=1}^n h_c\left(\sum_{j=1}^k \left\{X_i^{(j)} - k^{-1}T(X_i)\right\}^2 + w_1(\theta)\right) = 0. \tag{21}$$

There are two extreme cases, $c = \infty$ and $c = 0$. Clearly when $c = \infty$, $\hat{\theta}_n$ becomes the efficient estimate; and the efficient estimate of the variance σ^2 is equal to

$$-(2\hat{\theta}_n)^{-1} = \frac{1}{n(k-1)}\sum_{i=1}^n\sum_{j=1}^k \{X_i^{(j)} - k^{-1}T(X_i)\}^2.$$

This agrees with the result in [BKRW]. Next we consider the case when $c = 0$. The corresponding influence function is called the *most robust influence function* because its bound is the lowest among all influence functions. The corresponding estimate is called the *most robust estimate.*

Lemma 5. *Define an estimate of θ through*

$$-(2\tilde{\theta}_n)^{-1} = C_{k-1}^{-1} median\left(\sum_{j=1}^{k} \left\{ X_i^{(j)} - k^{-1}T(X_i) \right\}^2 : i = 1, \dots, n \right)$$

with $C_{k-1} = median(\chi^2_{k-1})$. Then $\tilde{\theta}_n$ is the most robust estimate of θ.

Proof. The random variables $\{\sum_{j=1}^{k}(X_i^{(j)} - T(X_i)/k)^2 : i = 1, \dots, n\}$ are i.i.d. samples from the $-(2\theta)^{-1}\chi^2_{k-1}$ distribution, which is a purely parametric model. Using the standard theory of robust estimation in parametric models (Hampel et al. (1986)), we can show that $\tilde{\theta}_n$ is an asymptotically linear estimate of θ with influence function

$$\psi_0 = \text{Constant } \times \text{sgn}\left(\sum_{j=1}^{k} \left\{ X^{(j)} - k^{-1}T(X) \right\}^2 + w_1(\theta) \right).$$

However, ψ_0 corresponds to the case $c = 0$ in (21). This completes the proof.

Theorem 3. *For $0 < c < \infty$, let $\hat{\theta}_n$ solve (21). Then $\hat{\theta}_n$ is an optimal robust estimate of θ corresponding to bound $C_0 = \lambda^{-1}c$.*

Proof. Since $w_1(\theta)$ is continuous and is strictly monotonic, the standard theory of M-estimates (e.g., Hampel et al. (1986), or Fernholz (1983)) may be applied here to show that $\hat{\theta}_n$ is asymptotically linear, with optimal influence function ψ_0.

The above procedure may be viewed as applying robust estimation theory to the conditional density of X given $T(X)$, which is a parametric model. This observation has enabled us to avoid estimating the marginal density of T.

Model (L). Note from (2) that for this model $T(x, \theta) = \sum_{j=1}^{k}(x^{(j)} - \theta)^2$. The score function is $\dot{\mathbf{l}}_1 = -2\sum_{j=1}^{k}(x^{(j)} - \theta)E(\eta|T)$. For $c > 0$, the optimal influence function has the form: $\psi_c = \lambda^{-1}h_c(\dot{\mathbf{l}}_1)$ for some constant λ, since by symmetry,

$$E\left(h_c \left\{ -2\sum_{j=1}^{k}(x^{(j)} - \theta)E(\eta|T) \right\} \middle| \sum_{j=1}^{k}(x^{(j)} - \theta)^2 \right) = 0.$$

For $c = 0$, we are led to $\psi_0 = \lambda^{-1}\text{sgn}(k^{-1}\sum_{j=1}^{k} x^{(j)} - \theta)$, the most robust influence function. The most robust estimate $\tilde{\theta}_n$ is the median of $\bar{X}_1, \dots, \bar{X}_n$, where $\bar{X}_i$

is the mean of $(X_i^{(1)}, \ldots, X_i^{(k)})$. For $c > 0$, one needs to estimate $E(\eta|T)$, which can be derived from the estimate of $p_T'(\cdot)/p_T(\cdot)$. Then the one-step procedure as defined in the last section gives the optimal estimate.

Two sample exponential mixture model. Recall that the distribution of the bivariate random vector $X = (X^{(1)}, X^{(2)})$ has density

$$f(X, \theta, G) = \int \exp\{\eta T(X, \theta) - b(\theta, \eta)\} dG(\eta),$$

where $T(X, \theta) = -(X^{(1)} + \theta X^{(2)})$ and $b(\theta, \eta) = -\ln(\eta^2 \theta)$. The marginal density of T is $p_T(t, G) = -$ constant $\times\, t \int \eta^2 \exp(\eta t) dG(\eta)$, for $t < 0$. The conditional density of $X^{(2)}$ given T is uniform $(0, -\theta^{-1}T)$. By symmetry, the efficient score function is

$$\begin{aligned} \dot{\mathbf{l}}^*(X, \nu, G) &= \{-X^{(2)} + E(X^{(2)}|T)\} E(\eta|T) \\ &= -\Big(X^{(2)} + \frac{T}{2\theta}\Big) E(\eta|T) \\ &= \Big(\frac{X^{(1)} - \theta X^{(2)}}{2\theta}\Big) E(\eta|T). \end{aligned}$$

The conditional expectation in the above formula is $E(\eta|T) = p_T'/p_T - T^{-1}$. To obtain the optimal influence function, we solve for w from $E\{h_c(-X^{(2)}E(\eta|T) + w)|T\} = 0$. By the symmetry of the uniform distribution again, $w(T) = -(T/2\theta)$ $E(\eta|T)$. Then the optimal influence function is given by

$$\psi_c(x) = \lambda^{-1} h_c \left\{ \frac{X^{(1)} - \theta X^{(2)}}{2\theta} \times E(\eta|T) \right\},$$

where λ is the normalizing constant. Thus, the optimal influence function may be expressed as $\psi(X) = \lambda^{-1} h_c(\dot{\mathbf{l}}^*)$. When $c = 0$, it becomes the most robust influence function: $\psi_0 = \lambda' \mathrm{sgn}(X^{(1)} - \theta X^{(2)})$. Note that the conditional expectation $E(\eta|T)$ disappears from the formula of ψ because it is always non-negative.

Lemma 6. *Suppose $X_i = (X_i^{(1)}, X_i^{(2)})$, $i = 1, \ldots, n$, are i.i.d. samples from the two sample exponential mixture model. Let $\tilde{\theta}_n = median\{X_i^{(1)}/X_i^{(2)} : i = 1, \ldots, n\}$. Then $\tilde{\theta}_n$ is the most robust estimate of θ.*

Proof. The proof is similar to that of Lemma 5.

Having obtained the preliminary estimate $\tilde{\theta}_n$, one can apply the procedure described in the last section to calculate the optimal estimate for any $c > 0$.

5. Simulation

We have conducted a small simulation study for the two sample exponential mixture model. In the models we have used, the true value for θ is equal to 1. The nuisance parameter η may come from the following mixing distributions:

1. Degenerate distribution: The distribution of η degenerates to $\eta \equiv 1$.
2. G = Uniform $(0.3, 6)$: The two sample exponentials are mixed over the uniform distribution.
3. G = Abs-normal $(4, 2)$: The two sample exponentials are mixed over the absolute values of $N(4, 2)$-distribution.
4. G = Log-normal $(1.5, 2)$: The two sample exponentials are mixed over the log-normal distribution.

Model 1 is in fact a parametric model with the main parameter θ and the nuisance parameter η. Models 2, 3, and 4 are genuine mixture models. Let each of the above models be denoted by $P_{\theta,G}$. The sample size in the simulation is $n = 100$. For each of the above models, the simulated data are generated from the contaminated distribution:

$$F_n = \left(1 - \frac{\delta}{\sqrt{n}}\right) P_{\theta,G} + \frac{\delta}{\sqrt{n}} H, \qquad \text{for } \delta \in \{0.0, 0.5, 1.0, 1.5\}. \tag{22}$$

In the above expression, the contaminating distribution H is chosen such that it causes the first variable $X^{(1)}$ to be of the form $X^{(1)} = \theta X^{(2)} + |N(6.5, 0.5)|$ and leaves the second variable $X^{(2)}$ unchanged. Since the ratio between $X^{(1)}$ and $X^{(2)}$ is of the order of θ, this contamination is chosen to significantly affect their relationship. For simplicity we do not put outliers in the distribution of $X^{(2)}$.

In each case, we use the most robust estimate $\tilde{\theta}_n$ as our initial estimate. Then we may calculate $T_i = -(X_i^{(1)} + \tilde{\theta}_n X_i^{(2)})$, for $i = 1, \ldots, 100$. Estimate the marginal density of T by

$$\hat{p}_n(t) = \frac{1}{n\sigma_n} \sum_{i=1}^{n} \left\{ \phi\left(\frac{t - T_i}{\sigma_n}\right) - \phi\left(\frac{t + T_i}{\sigma_n}\right) \right\}. \tag{23}$$

The bandwidth σ_n is chosen according to a similar rule in Silverman (1986): $\sigma_n = .79 R_0 n^{-\frac{1}{5}}$, where R_0 is the interquartile range of $T_1, \ldots, T_n$. The reflection in (23) is introduced to ensure that $\hat{p}_n(0) = 0$, which is the case for the true marginal density $p_T(t)$. Let $\hat{p}'_n(t)$ be the derivative of $\hat{p}_n(t)$. We can estimate the conditional expectation $E(\eta|T)$ by

$$\hat{E}(\eta|T = t) = \frac{\hat{p}'_n}{\hat{p}_n}(t) - \frac{1}{t}, \text{ for } t < 0.$$

For $c > 0$, define

$$\hat{\psi}(x, c) = \hat{\lambda}_n(c)^{-1} h_c \left\{ \frac{X^{(1)} - \tilde{\theta}_n X^{(2)}}{2\tilde{\theta}_n} \hat{E}(\eta|T) \right\}, \tag{24}$$

where $\hat{\lambda}_n(c)$ is defined in the same way as (13) using the whole data set ($m_n = 0$).

It is well known (e.g., Bickel (1981)) that when the underlying distribution is (22), the empirical version of the *asymptotic mean squared error* takes the form

$$\frac{(\delta c)^2}{\hat{\lambda}_n^2(c)} + \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_c^2(X_i), \tag{25}$$

where the first term estimates the squared bias and the second term estimates the variance. In the simulation we choose numerically the trimming parameter c to minimize (25). However, the true value of δ in (22) is usually unknown and has to be chosen subjectively. In the simulation we use $\delta_0 = 0.5$ to enter (25), no matter what δ is actually used to generate the data. Assume that the minimum of (25) is reached at $\hat{c}$. Then $\hat{\psi}(x, \hat{c})$ is used to construct the optimal estimate:

$$\hat{\theta}_n = \tilde{\theta}_n + \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_{\hat{c}}(X_i). \tag{26}$$

In our simulation we compare the performance of the following estimates:

(i) The parametric estimate: $\sum_{i=1}^{n} X_i^{(1)} / \sum_{i=1}^{n} X_i^{(2)}$ (see Lindsay (1983)).

(ii) The M-estimate: the solution of $\sum_{i=1}^{n} h_1(X_i^{(1)} - \theta X_i^{(2)}) = 0$.

(iii) The efficient estimate: the estimate defined by (24) with trimming parameter $c = 9.0$.

(iv) The optimal estimate $\hat{\theta}_n$: the estimate defined by (26).

(v) The most robust estimate: $\tilde{\theta}_n = median\{X_i^{(1)}/X_i^{(2)} : i = 1, \ldots, n\}$.

The simulation results are entered into Table 1. Each simulation is repeated 10,000 times. For each particular estimate, we calculate the Root Mean Squared Error by

$$\text{RMSE} = \left(\frac{1}{10000} \sum_{i=1}^{10000} 100 \times (\hat{\theta}_i - \theta)^2 \right)^{\frac{1}{2}},$$

where $\hat{\theta}_i$ is the estimate calculated from the ith sample. The figures in Table 1 are the ratios between the RMSEs of the estimates (i)-(iv) and those of the most robust estimates.

The standard errors are also calculated for the original RMSEs. With $10,000$ simulations, the standard errors are in the range from 0.005 to 0.02 for the situation without contamination. Therefore the numbers in Table 1 are accurate almost up to the second decimal points.

Table 1. Ratios of RMSEs

		δ			
Models	Estimates	0	0.5	1.0	1.5
Model 1	Parametric est.	0.70	1.45	1.76	1.73
	M-estimate	0.82	0.89	0.97	1.01
	Efficient est.	0.78	0.97	1.12	1.23
	Optimal est.	0.81	0.82	0.91	1.05
Model 2	Parametric est.	1.00	2.74	3.38	3.35
	M-estimate	0.81	1.07	1.26	1.36
	Efficient est.	0.92	0.97	1.05	1.26
	Optimal est.	0.91	0.79	0.79	1.07
Model 3	Parametric est.	3.64	4.10	4.57	4.30
	M-estimate	0.81	1.15	1.42	1.54
	Efficient est.	0.89	0.92	0.97	1.16
	Optimal est.	0.88	0.75	0.72	0.94
Model 4	Parametric est.	3.06	2.69	2.39	2.09
	M-estimate	1.06	1.39	1.69	1.85
	Efficient est.	1.12	1.24	1.47	1.71
	Optimal est.	1.11	1.12	1.36	1.71

Conclusion

From Table 1 one can see that the parametric estimate is inferior to other estimates except in the parametric model where it is supposed to be the best. The parametric estimate is also quite sensitive to contamination.

The second finding is that even at the true model ($\delta = 0$), the optimal estimate is no worse than the efficient estimate. This is a little surprising however. One possible explanation is that the exponential distribution has relatively long tails and we may have benefited from using the optimal estimates. Another possible explanation may be that we need better estimates of the score function to improve the performance of the efficient estimates. It is also interesting to note that the M-estimate behaves well for this situation. Thus the M-estimate may serve as a simple estimate when the chance for contamination is small. It is encouraging to see that in most models, as the amount of contamination increases, the optimal estimates tend to out-perform the other estimates. This agrees with the theoretical results.

We have paid special attention to the most robust estimates as they are easy to compute. When the contamination is moderate, the most robust estimates have relatively large RMSEs (the ratios < 1). However, as the amount of contamination increases, the most robust estimates start to show better performance (the ratios ≥ 1).

Acknowledgement

This work is part of the author's Ph.D. dissertation that was carried out at the University of California, Berkeley, under the supervision of Professor P. J. Bickel. The author would like to thank him for many helpful discussions leading to this paper. The author is also indebted to the Associate Editor and an anonymous referee for their comments.

References

Bickel, P. J. (1981). Quelques aspects de la statistique robuste. $LectureNotesinMath\triangleright$ **876**. Springer-Verlag, Berlin.

Bickel, P. J. (1982). On adaptive estimation. Ann. Statist. **10**, 647-671.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). Efficient and Adaptive Estimation in Semiparametric Models. Johns Hopkins University Press.

Fernholz, L. T. (1983). Von Mises' Calculus for Statistical Functionals. Springer-Verlag.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust Statistics: The approach Based on Influence Functions. John Wiley, New York.

Klaassen, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. Ann. Statist. **15**, 1548-1562.

Lindsay, B. G. (1983). Efficiency of the conditional score in a mixture setting. Ann. Statist. **11**, 486-497.

LeCam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. Proc. 3rd Berkeley Symp. Math. Statist. Probab. **1**, 129-156.

Neyman, E. and Scott, E. (1948). Consistent estimates based on partially consistent observations. Econometrica **16**, 1-32.

Shen, L. Z. (1992). Robust estimation in semiparametric models. Ph.D. Dissertation, University of California, Berkeley.

Shen, L. Z. (1994). On optimal B-robust influence functions in multidimensional parametric models. Comm. Statist. Theory Methods **23**, 1103-1122.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall.

van der Vaart, A. W. (1988). Estimating a real parameter in a class of semiparametric models. Ann. Statist. **16**, 1450-1474.

Wu, C. O. (1990). Asymptotically efficient robust estimates in some semiparametric models. Ph.D. Dissertation, University of California, Berkeley.

Biometrics and Statistical Sciences, The Procter and Gamble Company, Cincinnati, OH 45242, U.S.A.

(Received February 1993; accepted May 1994)